Attorney Docket No.: 016757-000314US

PATENT APPLICATION

METHOD OF PERFORMING SPEECH RECOGNITION ACROSS A NETWORK

Inventor(s): Todd F. Mozer, a citizen of The United States, residing at
24275 Elise Court
Los Altos Hills, CA 94024

Forrest S. Mozer, a citizen of The United States, residing at
38 Somerset Place
Berkeley, CA 94707

Assignee: Sensory, Incorporated
1991 Russell Avenue,
Santa Clara, CA, 95054

Entity: Small

TOWNSEND and TOWNSEND and CREW LLP
Two Embarcadero Center, 8th Floor
San Francisco, California 94111-3834
Tel: 650-326-2400

METHOD OF PERFORMING SPEECH RECOGNITION ACROSS A NETWORK

CROSS-REFERENCES TO RELATED APPLICATIONS

[0001] This application is a continuation of and claims the benefit of U.S. Patent Application Serial No. 10/051,838, filed January 16, 2002, which is continuation of and claims the benefit of U.S. Patent Application Serial No. 09/328,656, filed June 9, 1999, which is a continuation-in-part of and claims the benefit of U.S. Patent Application No. 08/822,852, filed March 24, 1997, which claims priority from U.S. Provisional Application Serial No. 60/032,788, filed December 6, 1996. The 10/051,838, 09/328,656, 08/822,852, and 60/032,788 are hereby incorporated herein by reference.

BACKGROUND OF THE INVENTION

[0002] The present invention relates to speech recognition and more particularly to inexpensive and user friendly speech recognition techniques.

[0003] Speech recognition has been extensively studied for several decades because of its interest on intellectual grounds and because of its military and commercial applications. Some of the commercial applications involve speaker verification and improving the man-machine interface (*e.g.*, U.S. Patent Nos. 3,742,143; 4,049,913; 4,882,685; 5,281,143; and 5,297,183). As evidence of the extensive research on speech recognition, the U.S. Patent Office has granted more than 600 patents on speech recognition or related topics in the last three decades and as many as 10,000 articles have appeared in the scientific or engineering literature during that time.

[0004] Generally, a speech recognition device analyzes an unknown audio signal to generate a pattern that contains the acoustically significant information in the utterance. This information typically includes the audio signal power in several frequency bands and the important frequencies in the waveform, each as a function of time. The power may be obtained through the use of bandpass filters (*e.g.*, U.S. Pat. No. 5,285,552) or fast Fourier transforms (*i.e.*, FFTs) (*e.g.*, U.S. Pat. No. 5,313,531). The frequency information may be obtained from the FFTs or by counting zero crossings in the filtered input waveform (U.S. Pat. No. 4,388,495).

[0005] Speech recognition devices can be classified as "speaker dependent" or "speaker independent." Speaker dependent devices require that the user train the system by speaking all of the utterances in the entire recognition set several times. Speaker independent devices do not require such training because the acoustic cues obtained from many repetitions of the utterances in the recognition set, as spoken by many different speakers, are used to train the recognizer to recognize an unknown utterance by a speaker whose phrase was not part of the training set.

[0006] Commercial applications of both speaker independent and speaker dependent recognition are becoming prevalent for applications such as voice activated phone dialing, computer command and control, telephone inquiries, voice recorders, electronic learning aids, data entry, menu selection, and data base searching. The growth of the speech recognition marketplace results from the decreasing cost of computing power and recognition technology as well as the need for more friendly user interfaces.

[0007] In some applications, speaker dependent recognition is required because the user must input information that he/she later requests. An example is voice dialing, which is being test marketed by U.S. West among others, in which the user verbally enters a directory of names and phone numbers. This information is later solicited by using speaker dependent recognition when the user wishes to make a phone call. Except for applications such as voice dialing that require speaker dependent recognition, this technology has not achieved wide market acceptance because it is not user-friendly due to the required training.

[0008] Much of the interest in speaker independent recognition is because of the simpler user interface. An example of a speaker independent recognition software package running on personal computers is VOICE Release 2.0 from Kurzweil AI, which is able to recognize as many as 60,000 words without user training. Other examples of similar technologies are the IBM Voice Type 3.0, used in radiology, the Wild Card LawTALK, used in legal applications, and the Cortex Medical Management, used for anatomic pathology. More than two dozen speaker independent recognition computer products are available and they all require considerable computing power to perform the sophisticated natural language processing involving context, semantics, phonetics, prosody, *etc.*, that is required to recognize very large sets of utterances without user training. Hence, large vocabulary, speaker independent recognition products require considerable computing power.

[0009] Small vocabulary, speaker independent recognition also appears in commercial applications where the number of utterances to be recognized is limited. Examples are the Sensory, Inc. speaker independent recognition LSI chip (U.S. Pat. No. 5,790,754) used in electronic learning aids such as the Fisher-Price Radar product, or in time setting applications such as the VoiceIt clock. This technology is accurate and inexpensive but, in the current art, it is limited to use with relatively small vocabularies because the LSI chip does not contain the computing power required for natural language processing or the memory required to store information about a very large inventory of recognition words.

[0010] The above described limitations of current recognition technology narrow the range of its applicability in consumer electronic products. For example, it would be desirable to select a particular song from a compact disk changer that holds many compact disks by telling it which disk and which song on that disk you wish to hear. This is not currently feasible because solving this problem with speaker dependent recognition requires that the user repeat the names of all recordings on every compact disk that he owns, while solving it with speaker independent technology would require that the recognizer be able to understand the name of every song on every compact disk in the world. Or, consider the use of speech recognition during the interaction of a surfer with an internet website. Most of this interaction is at a simple one-step-at-a-time level where the vocabulary to be recognized at each step is small but the total vocabulary associated with all of the steps may be large. For this application, speaker dependent recognition may not be feasible because of its inconvenience. Speaker independent recognition is feasible, but, in the current art, analyzing the speech by the web site's main processor creates conflicts between the recognition program and the application and may slow down the application to the point that use of recognition becomes unacceptable to the user. Also, adding additional processing power to handle the speaker independent recognition may not be feasible due to its cost.

SUMMARY

[0011] The present invention provides an inexpensive and user-friendly speaker independent speech recognition system. A speech recognition system according to the present invention may function without the use of natural language processing or internal storage of large amounts of speech recognition data.

[0012] In one embodiment, an inexpensive, speaker independent recognition engine is placed in the base unit of an electronic apparatus. Depending on the application, the base unit may be a compact disk player, computer, internet access device, video game player, television set, telephone, *etc.* The recognition engine may be a software program running in a general purpose microprocessor or an LSI chip such as the Sensory RSC-164 available from the assignee of the present application. Since the recognition engine should be inexpensive, it may be capable of recognizing only a limited set of utterances at any one time, although this recognition set of utterances may change from one application of recognition to the next in the same base unit.

[0013] The architecture of the product is such that, in operation, an external medium is connected to the base unit. The external medium may be a compact disk if the base unit is a compact disk changer, a floppy disk if the base unit is a computer, a video game cartridge if the base unit is a video game player, a cable or rf transmission if the base unit is a television set or an internet access device, a phone cable if the base unit is a telephone, *etc.* Included in the information provided to the base unit by the external medium is the data required for the recognition engine to recognize a spoken utterance from a limited set of candidate utterances. As the interaction between the base unit and the user progresses, different sets of data may be supplied by the external medium to the recognition engine in the base unit in order to allow different recognition sets at different times in the interaction.

[0014] Or, in some applications, only one or two data sets might ever be supplied from the external medium to the base unit. Consider the case of a watch that utilizes speech recognition for setting the time. To function, this watch might require two speaker-independent recognition sets, the first of which would be the digits, and the second of which would be the words "set," "hours," "minutes," "seconds," and "done." A problem is that worldwide sales require that this watch perform speech recognition in any of dozens of languages. In the current art, this would require either that the watch manufacturer and retailers carry inventories of a large number of different units or that the watch is loaded with information in many languages, at an unacceptable expense. An alternative approach would be to include a small amount of programmable, non-volatile memory in the watch, and to download, from the Internet, the pertinent information for whatever language a purchaser wishes his watch to recognize. The voice prompts required to guide the user through setting the time would also be downloaded in the language of the user's choice in the same way.

Downloading information to devices from the Internet is already a normal operation and watches with infra-red interfaces to computers are available in the market.

[0015] In accordance with a first aspect of the present invention, a base unit is provided wherein features of spoken utterances are analyzed by a programmable pattern recognition system to provide recognition results. A method of operating the base unit includes steps of programming the pattern recognition system to recognize a first set of words, operating the pattern recognition system as programmed to generate at least a first recognition result responsive to input speech, retrieving programming information for the pattern recognition system from a source external to the base unit responsive to the first recognition result and reprogramming the pattern recognition system to recognize a second set of words selected responsive to the first recognition result.

[0016] In accordance with a second aspect for he present invention, a method for speaker-independent speech recognition includes steps of performing speaker-independent speech recognition of user utterances in a base unit, receiving, in the base unit, first information pertinent to the speech recognition from an external medium, and receiving, in the base unit, second information independent from the first information and related to the user utterances from the external medium.

[0017] In accordance with a third aspect of the present invention, a method for speaker-independent speech recognition includes steps of downloading from an external medium into a base unit the information required for the speech recognition to operate in a selected one or a few of several different languages.

[0018] A further understanding of the nature and advantages of the inventions here may be realized by reference to the remaining portions of the specification and the attached drawings.

BRIEF DESCRIPTION OF THE DRAWINGS

[0019] Fig. 1 depicts a general representation of an apparatus incorporating speech recognition according to one embodiment of the present invention;

[0020] Fig. 2 is a flowchart describing steps of the operation of the apparatus of Fig. 1 in accordance with one embodiment of the present invention;

[0021] Fig. 3 depicts a particular example of the apparatus of Fig. 1, a compact disk changer enhanced in accordance with one embodiment of the present invention; and

[0022] Fig. 4 depicts the operation of the compact disk changer of Fig. 3 in accordance with one embodiment of the present invention.

DETAILED DESCRIPTION

[0023] Fig. 1 depicts a general representation of an apparatus 100 incorporating speech recognition according to one embodiment of the present invention. Apparatus 100 includes a base unit 102 and an external medium 104. Apparatus 100 may provide speech recognition capabilities to, for example, various electronic appliances such as a compact disk changer, telephone, computer, television, watch, *etc.* Components of apparatus 100 may perform other functions besides speech recognition in the context of such appliances.

[0024] Base unit 100 includes a microphone 105, a feature extraction unit 106, a programmable pattern recognition system 108, a weight memory 110, a pattern recognition programmer 112, a user interface system 113, a speaker 114, a graphical display 116, and an external interface 118. It is to be understood that not all of these elements are required for any particular embodiment of the present invention. Also, many of the depicted elements are implementable in either hardware or software.

[0025] Microphone 105 accepts user speech utterances and converts them to an analog electrical signal. Feature extraction unit 106 converts the analog electrical signal to digital information and extracts features which characterize the input utterances to facilitate recognition. Feature extraction unit 106 may be implemented in any one of a number of ways in either hardware or software. One preferred implementation of feature extraction unit 106 is taught in co-assigned U.S. Pat. No. 5,790,754, the contents of which are herein incorporated by references for all purposes.

[0026] Pattern recognition system 108 recognizes the spoken utterances based on the features extracted by feature extraction unit 106. Preferably, pattern recognition system 108 is a neural network that employs weights from weight memory 110. An example of such a neural network is found in U.S. Pat. No. 5,790,754. Pattern recognition system 108 selects a recognition result for the input utterance from among members of a presently selected recognition set.

[0027] The weights have been previously developed by training the neural network with multiple examples of the specific utterances comprising the recognition set associated with these weights. For example, if the recognition set consists of the words "yes" and "no," several hundred examples of each of these words, collected from the same population that will use the product, might be used to train the neural network. Another set of weights might be associated with the digits from 0 through 9.

[0028] In accordance with the present invention, the recognition set and associated weight set may change over time. Thus, at a specific time in an application, pattern recognition system 108 might distinguish yes from no by using one weight set and, at another time, it might recognize the single digits by using the second weight set. In this way, a large number of different utterances can be recognized without any one recognition set being so large that a more sophisticated recognition engine is required.

[0029] Pattern recognition programming system 112 controls the selection of a current recognition set and weight set, at least partially in response to the recognition results generated by pattern recognition system 108. The user interface system presents output to the user through speaker 104 and/or graphical display 116. The information presented to the user may include prompts for input to microphone 105 or application specific information. User interface system 113 may incorporate a speech synthesis capability.

[0030] Pattern recognition programming system 112 employs external interface 118 to retrieve new recognition sets and weight sets into weight memory 110. External interface 118 may be a storage interface, *e.g.*, an IDE or SCSI interface, a network interface as would be used with a local network, or a network interface to an internet or intranet. External interface 118 may also be modem for connection to a telephone line, a modem for connecting to a CATV network, or a wireless modem for sending and receiving electromagnetic transmissions. External medium 104 may be, *e.g.*, a compact disk, a compact disk jukebox, a remote server, a web site, a floppy disk, a hard drive, a video game cartridge, *etc.* The connection between external interface 118 and external medium 104 may be a SCSI port, an IDE port, a telephone line, an intranet, the Internet, a CATV network, the airwaves, *etc.*

[0031] Software or computer code to implement any of the elements of Fig. 1 may be stored in, for example, a memory device, CD-ROM, floppy disk, hard drive, any computer-readable storage medium, *etc.*

[0032] Fig. 2 is a flowchart describing steps of the operation of the apparatus of Fig. 1 in accordance with one embodiment of the present invention. At step 202, pattern recognition programming system 112 accesses external medium 104 to verify it in fact contains recognition set data and weight set data of the kind employed by pattern recognition system

108 and to determine the number of recognition sets and associated weight sets stored there. At step 204, pattern recognition programming system 212 retrieves an initial set of words and associated weight set into weight memory 110. At step 206, microphone 105 picks up a user's speech utterance. Feature extraction unit 106 develops a set of features to characterize the user utterance. Pattern recognition system 108 recognizes the utterance based on the weights currently stored on weight memory 110. The utterance may come in response to a prompt conveyed to the user by user interface system 113 via speaker 114 or graphical display 116.

[0033] At step 208, pattern recognition programming system 112 receives the recognition result and selects a new set of words and associated weight set based on this result. The new recognition set and weight set are transferred from external medium 104 to weight memory 110 through external interface 118. In some embodiments, other information, independent from the recognition set and weight set information, is also retrieved from external medium 104. At step 210, user interface system 113 presents this other information or the result of processing this other information to the user. For example, the information may be audio data, and user interface system 113 may play a song. Alternatively, the information may be video data and user interface system 113 may display an image, video program, or scene from a video game. At step 212, base unit 102 receives and recognizes a new user utterance but using the newly loaded recognition set and weight set information.

[0034] Steps 208, 210, and 212 repeat as often as required by the application. It is of course not necessary that new recognition weight set information be loaded after each utterance that is recognized. For example, the watch that sets time by use of speaker-independent recognition in any of several languages might have only one or two weight sets, pertinent to a specific language, downloaded from external medium 109 through step 204, once during the life of the watch.

[0035] Fig. 3 depicts a particular example of the apparatus of Fig. 1, a compact disk changer enhanced in accordance with one embodiment of the present invention. A compact disk changer 300 incorporates the functionality of base unit 102 along with circuitry necessary for compact disk changer operation. An integrated circuit 302 includes feature extraction unit 106, pattern recognition system 108, and weight memory 110. Integrated circuit 302 is preferably the RSC-164 speech recognition LSI chip manufactured by Sensory, Inc., assignee of the present application. Integrated circuit 302 is also capable of synthesizing speech from stored data and this capability is utilized by the compact disk changer enhanced in accordance with the present invention. Attached to compact disk changer 300 is a "jukebox" 304, into which compact disks 306 may be loaded. Compact disks 306 perform the function of external medium 104. They store recognition weight data and other information in the form of audio data to be played.

[0036] The weight sets utilized by integrated circuit 302 are located in each of the compact disks 306. Jukebox 304 is assumed to be capable of storing as many as 24 compact disks and loading the selected disk for playing. Some of these 24 slots may be empty. For illustration, it is assumed that ten compact disks are in jukebox 304 and six of them are of the type that contain weight and recognition set information. Each of these six compact disks 306 contain weights for two sets of words, the first of which is the name of the compact disk and the second of which is the list of songs in that compact disk.

[0037] Fig. 4 depicts the operation of the compact disk changer of Fig. 3 in accordance with one embodiment of the present invention. After being turned on, at step 402, compact disk changer 300 scans external medium 304 and checks for appropriate signals from the six compact disks which indicate that they are of the type containing recognition weights. If the compact disks are of the correct type, compact disk changer 302 receives the required information on the number and type of weight sets in each of these compact disks at step 404. If the compact disks do not contain the weight set information, integrated circuit 302 synthesizes and outputs the spoken phrase "Please load manually" at step 406 to indicate that spoken control will not be possible.

[0038] At step 408, integrated circuit 302 then synthesizes and outputs the spoken phrase "Which compact disk should I load?" and it analyzes the audio response. The first weight set from each of the six compact disks 306 are downloaded into compact disk changer 300 and used by integrated circuit 302 to decide which of the compact disks was requested by the speaker at step 410. Suppose the compact disk with music by Montovani was selected. From that compact disk, compact disk changer 300 downloads speech data in compressed form at step 414 and generates "I will play CD Montovani." Also, at step 414, the Montovani compact disk is then loaded into the compact disk changer and integrated circuit 302 then generates the phrase "Which song should I play?" The second weight set on the Montovani compact disk is downloaded and used by integrated circuit 302 to determine which song was selected at step 416.

[0039] Compact disk changer 300 then downloads the appropriate audio data from the compact disk and plays this song at step 418 and repeats the above selection process by going to step 410. If a compact disk is requested that is not in jukebox 304 or if a song is requested that is not in the selected compact disk, integrated circuit 302 generates the phrase "Not available. Please load manually" at step 420.

[0040] Through use of the invention, a large number of utterances may be recognized by a relatively simple recognition engine because an over-large number of utterances is not contained in any recognition set. Furthermore, by use of the invention, devices that operate in a user-friendly manner are achieved because they require no training of the recognizer.

[0041] Another embodiment having features similar to those in the specific embodiment would be a computer that contains a recognition engine and that receives weight sets from software packages. In this way, the software manufacturer can add speech recognition to his word processor, spread sheet program, data base program, game, *etc*. For this application, external interface 118 operates to access a hard disk, CD-ROM, or floppy.

[0042] Similarly, Internet web sites can offer speech recognition by downloading weights, *e.g.*, in the form of Java applets, to the local computer. This offers new possibilities for interactions such as learning. For example, suppose a child selects a web site for learning more about numbers. The site can download recognition sets, speech data, and screen graphics to the child's computer, which then displays a farm scene that includes 5 chickens. The downloaded speech can then say "How many chickens are there in the picture?" The child answers "five." The recognition program decides the answer and feeds this information to the web site, which sends new recognition information, speech, and graphics back to the local computer in order to say "You're right!" and to continue the interaction. This type of interactive learning tool is especially beneficial for children whose natural response is speech, not interactions with a mouse, joystick, or keyboard. For this application, external interface 118 would operate as a network interface or modem in combination with the well-known protocols for accessing the Internet.

[0043] Cable TV or satellite television transmissions can include recognition weights that are used by the receiving TV set to select programs through speech recognition. Through this capability, one can also play video games on the television set using speech as both a game output and user input during the game play, with both input and output speech synchronized with graphics on the TV screen. For this application, external interface 118 operates as an RF receiver, receiving both recognition weight information and other video and/or audio information.

[0044] While the above are complete descriptions of preferred and other embodiments of the invention, other arrangements and equivalents are possible and may be employed without departing from the true spirit and scope of the invention. The terms and expressions which have been employed here are used as terms of description and not of limitations, and there is no intention, in the use of such terms and expressions, of excluding equivalents of the features shown and described, or portions thereof, it being recognized that various modifications are possible within the scope of the appended claims and their full scope of equivalents.